Supplemental Information

A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes

Ran Li, Mian Gong, Xinmiao Zhang, Fei Wang, Zhenyu Liu, Lei Zhang, Qimeng Yang, Yuan Xu, Mengsi Xu, Huanhuan Zhang, Yunfeng Zhang, Xuelei Dai, Yuanpeng Gao, Zhuangbiao Zhang, Wenwen Fang, Yuta Yang, Weiwei Fu, Chunna Cao, Peng Yang, Zeinab Amiri Ghanatsaman, Niloufar Jafarpour Negari, Hojjat Asadollahpour Nanaei, Xiangpeng Yue, Yuxuan Song, Xianyong Lan, Weidong Deng, Xihong Wang, Chuanying Pan, Ruidong Xiang, Eveline M. Ibeagha-Awemu, Pat (J.S.) Heslop-Harrison, Benjamin D. Rosen, Johannes A. Lenstra, Shangquan Gan, Yu Jiang

Correspondence: yu.jiang@nwafu.edu.cn; shangquangan@163.com

This PDF file includes:

Supplemental Methods

Genome size estimation

Based on the *k*-mer method, the sheep genome size was estimated using gce-1.0.2 (Liu et al. 2013). The Illumina clean paired-end reads from the HiFi sequencing animals were used. The 17-mer distribution showed a major peak depth at 18× to 29×. Based on the number of *k*-mers and relative *k*-mer depth for each animal, we estimated the genome size of sheep to be 2.84 to 3.10 Gb, according to the formula: Genome size = *k*-mer_number/Peak_depth.

Genome annotation of non-reference sequences

RepeatMasker v4.0.5 (http://www.repeatmasker.org) was used to softmask the non-reference sequences. Then we performed an ab initio gene structure prediction using AUGUSTUS v3.3.3 (Stanke et al. 2006) and searched the protein sequence predicted by AUGUSTUS against the local protein database using DIAMOND (Buchfink et al. 2021) BLASTP (parameters: --more-sensitive --evalue 1e-10 --max-target-seqs 1) and kept hits with minimum coverage of 70% and identity of 80%. The local protein database was built using DIAMOND (Buchfink et al. 2021) makedb with RefSeq protein sequences from sheep, goat and cattle.

Annotation of repeats

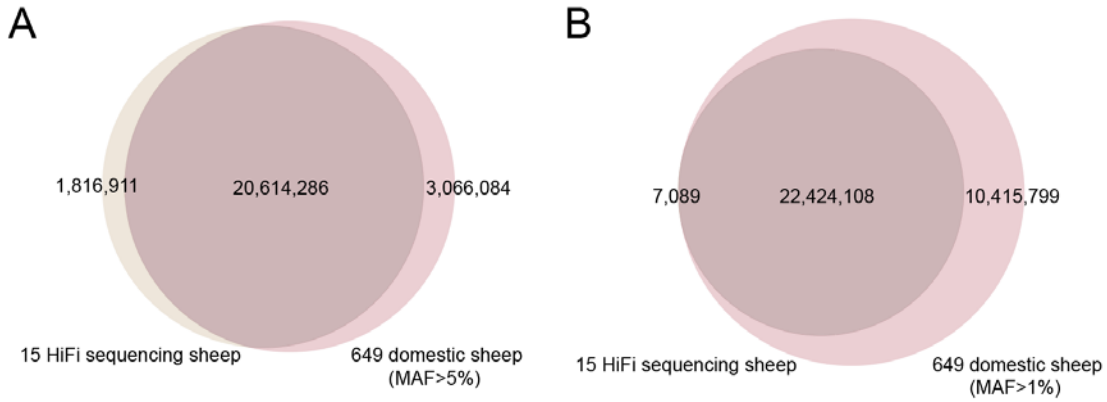
Interspersed repeats and low complexity DNA sequences were identified using RepeatMasker v4.0.5 (http://www.repeatmasker.org) with parameters: -species Ruminantia -xsmall -s -no_is -cutoff 255 -frag 20000 -gff. Tandem repeats were identified using Tandem Repeat Finder v4.09 (Benson 1999). The identified tandem repeats as well as those low complexity sequences identified by RepeatMasker were merged as low-complexity regions. Those with unit motif lengths $\geqslant$7 bp were classified as VNTR and with unit motif lengths $\leqslant$6 bp as STR. Sequence divergence rate of TE repeats were obtained from the RepeatMasker output.

Segmental duplications (SDs) were identified using BISER v1.2.3 (https://github.com/0xTCG/biser) on a masked version of the ARS-UI_Ramb_v2.0 assembly without unplaced scaffolds. The SDs identified by BISER were further filtered as described previously (Vollger et al. 2022): 1) ≤10% mismatch rate in the alignment, 2) ≤50% gap rate in the alignment, 3) ≥1 kb of aligned sequence, and 4) ≤70% satellite sequence as identified by RepeatMasker v4.0.5 (http://www.repeatmasker.org).

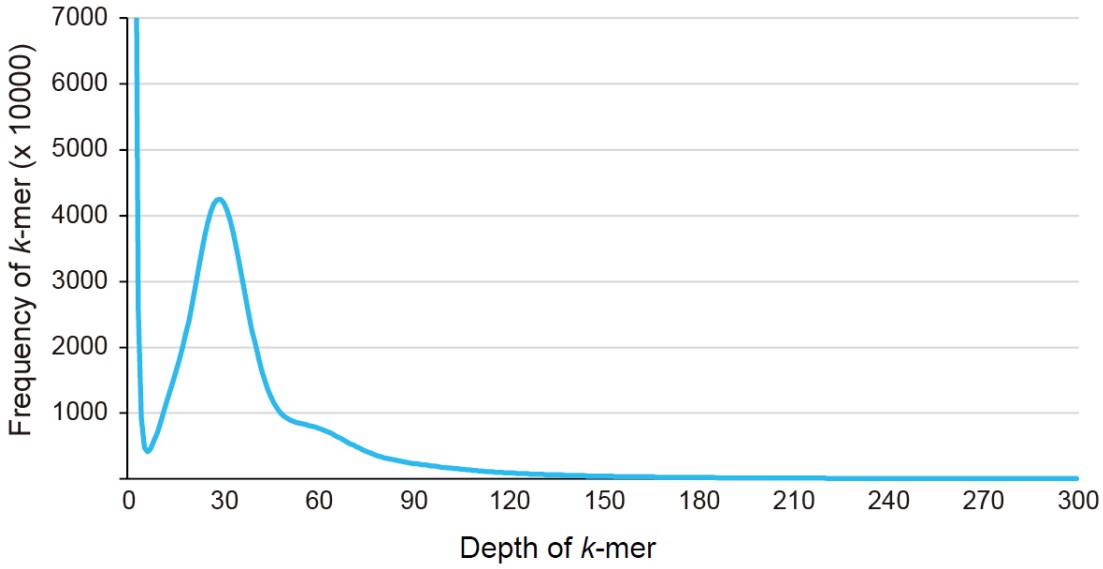
SNP calling from whole genome sequencing data

SNP calling was performed for the whole genome sequencing data as described in above section following our previous workflow (Chen et al. 2018). Briefly, clean reads were mapped to ARS-UI_Ramb_v2.0 using BWA-MEM v0.7.17 with default parameters. Duplicate reads were excluded using Picard MarkDuplicates (http://broadinstitute.github.io/picard/). Genome Analysis Toolkit (GATK v4.2.0.0) (McKenna et al. 2010) HaplotypeCaller module was used to generate gVCF file for each sample and VariantFiltration module was used to filter false calls "QD <2.0 || FS > 60.0 || MQRankSum <-12.5 || ReadPosRankSum < -8.0 || SOR >3.0 || MQ <40.0". SNPs were further filtered as below using VCFtools (Danecek et al. 2011): (1) biallelic variation; (2) missing rate <0.1; (3) mean reads depth (DP) >1/3× and <3× of total sequencing depth. We also excluded flanking SNPs within 100 bp of SV breakpoints to avoid the potential inaccurate SNPs near the SV regions.

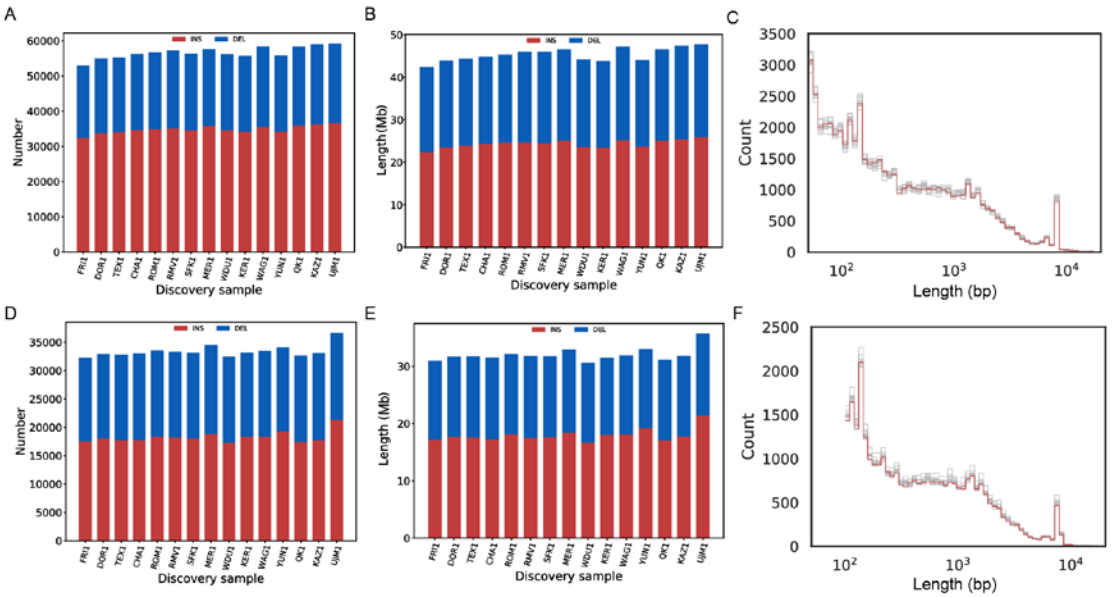
Supplemental Figures



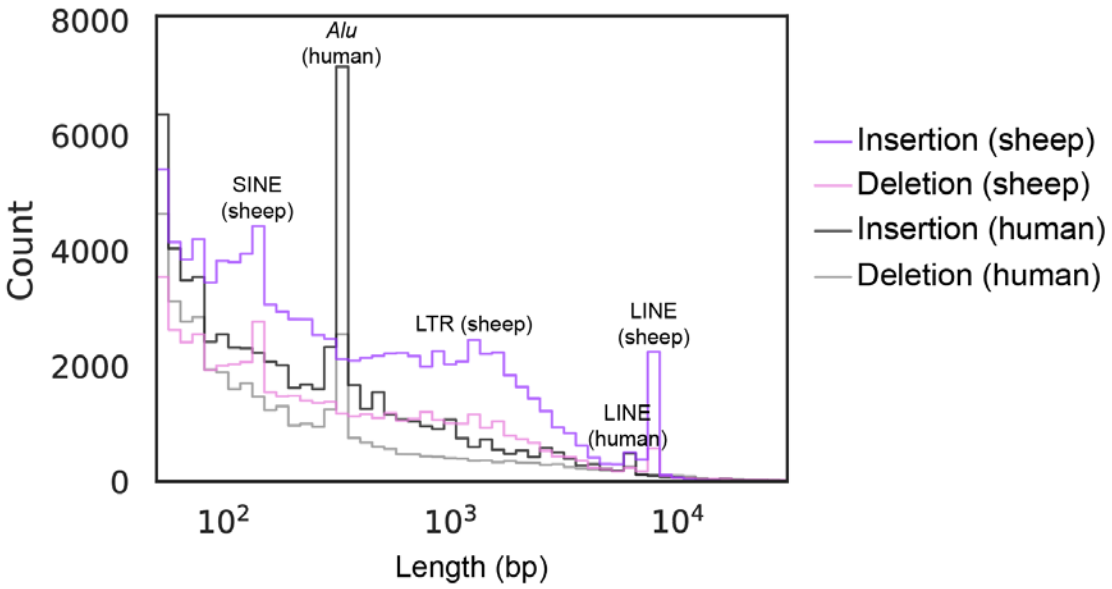
Supplemental Fig. S1 Venn diagram showing the number of SNPs from 649 domestic sheep of Illumina short reads sequencing with MAF>0.05 (A) and MAF>0.01 (B), comparing with all the SNPs from the 15 HiFi sequencing sheep.



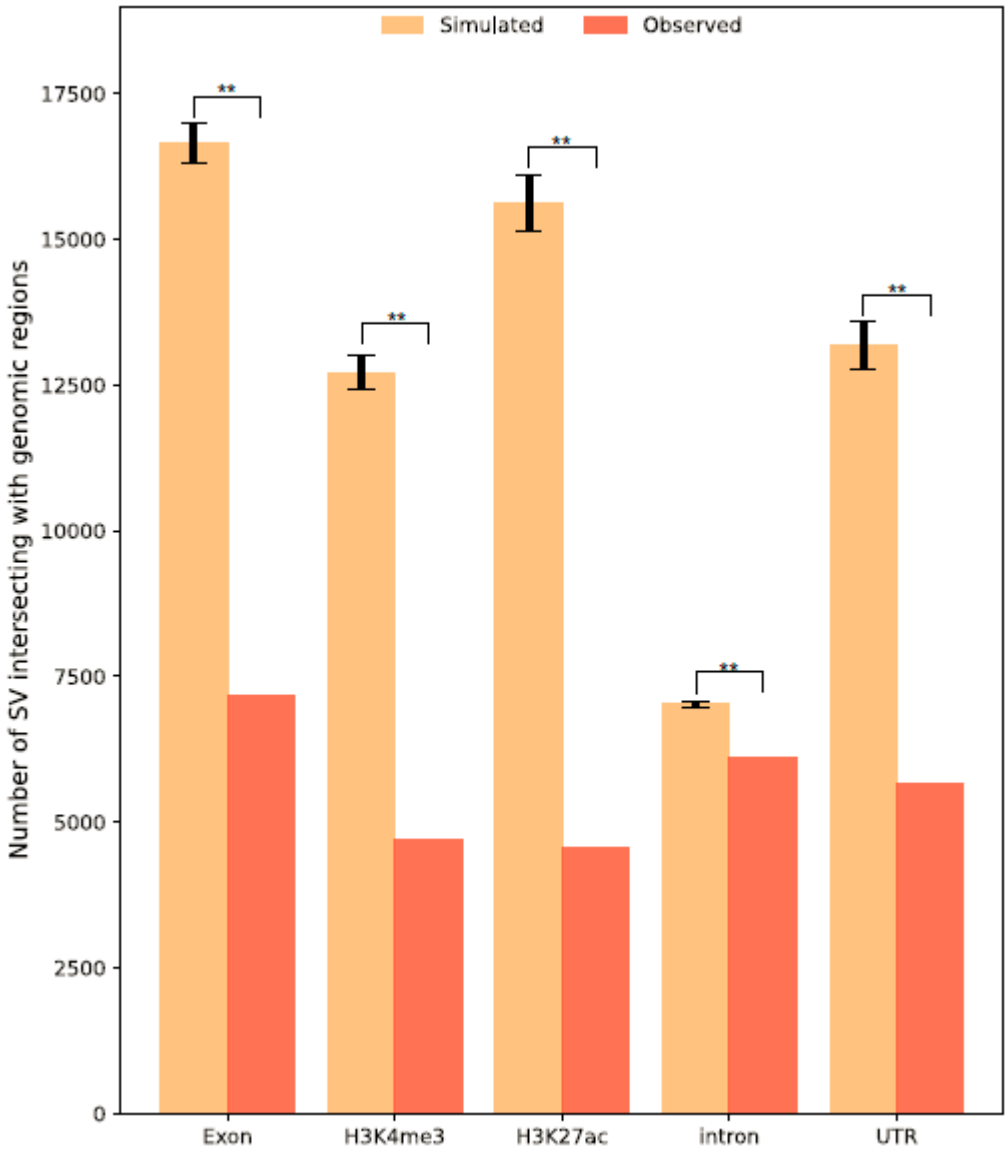
Supplemental Fig. S2 *K*-mer analysis for estimating the genome size of sheep. Totally 95 Gb Illumina clean paired-end reads from FRI1 were used. The 17-mer distribution showed a major peak at 29×. The genome size of sheep is estimated to be 2.84 Gb, according to the formula: Genome size = *k*-mer_number/Peak_depth.



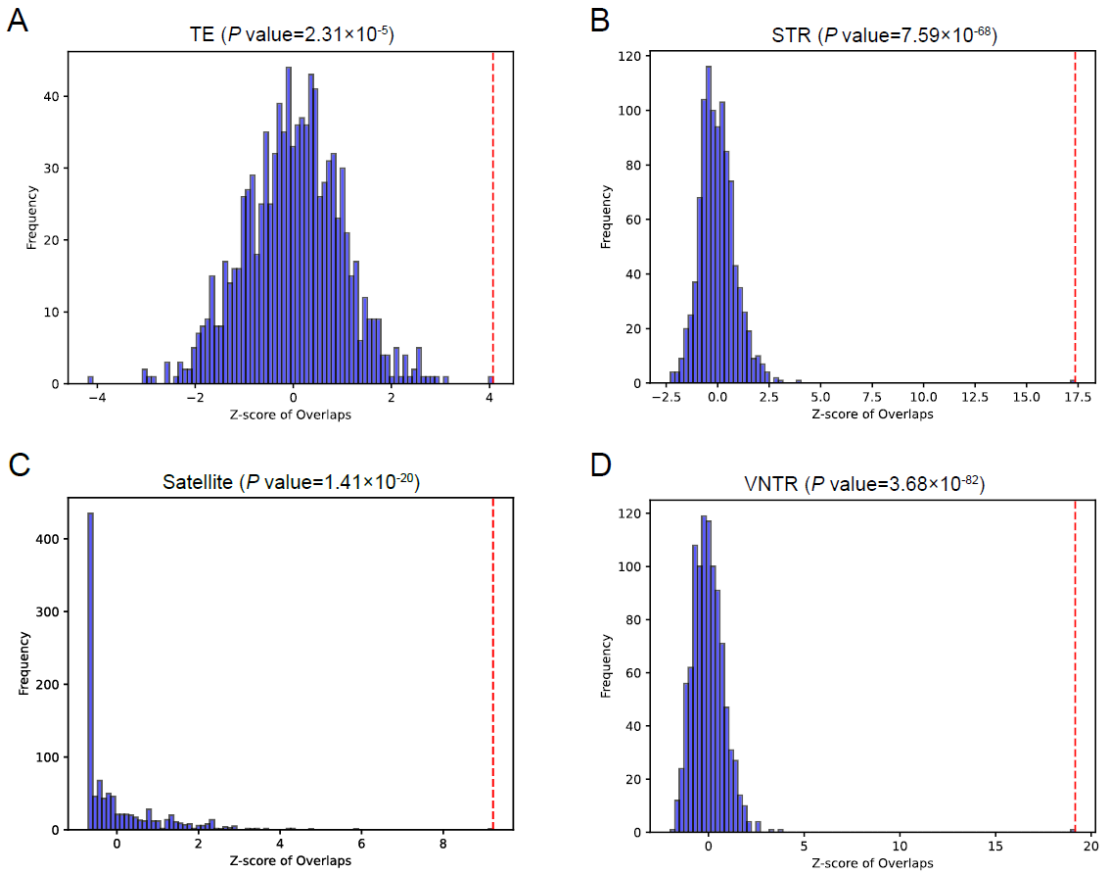
Supplemental Fig. S3 The number and length distribution of SVs detected from WDU1 as compared with other individuals. SVs were detected either by read-based using pbsv (A-C) or assembly-based approach using minigraph (D-F). In (C) and (F) for length distribution, the WDU1 is colored in red whereas the others in grey.



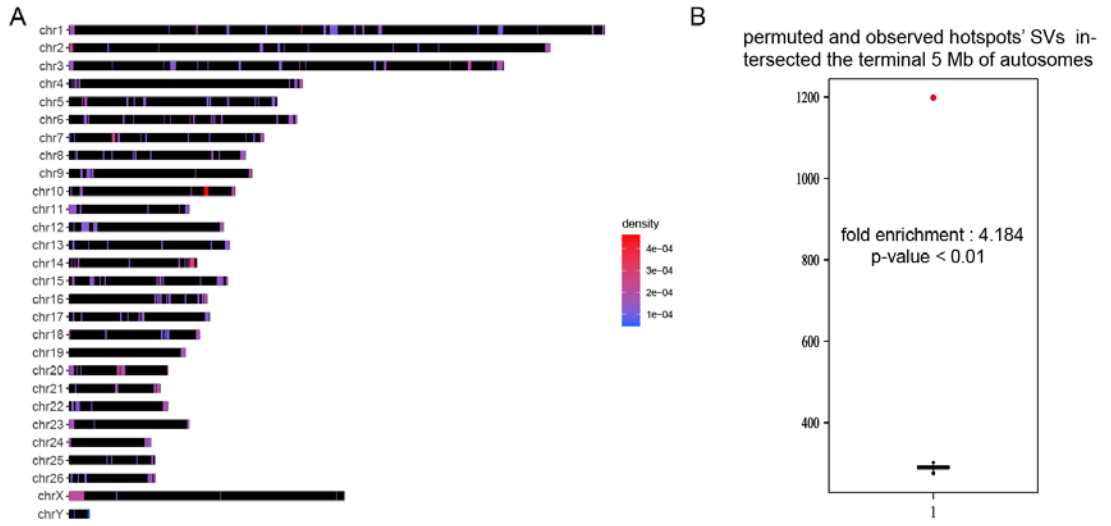
Supplemental Fig. S4 Length distribution of SVs in sheep as compared with humans.



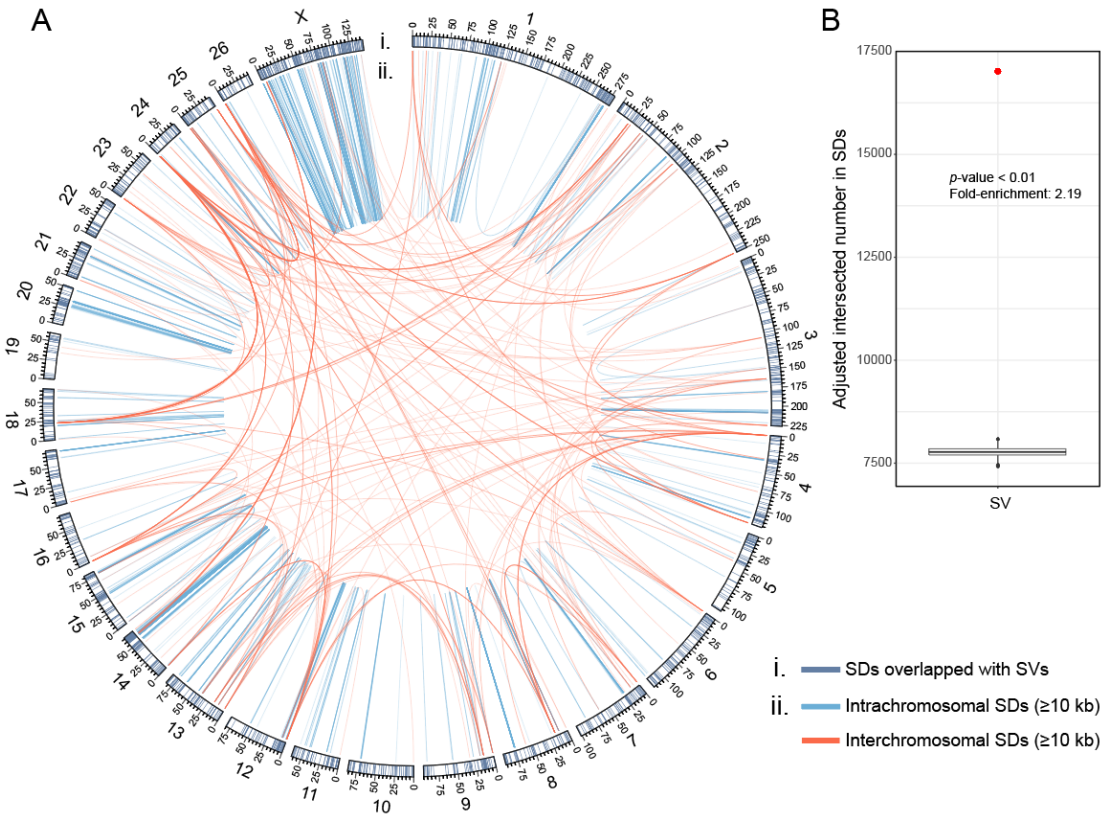
Supplemental Fig. S5 The number of SVs (right bar) intersecting functional elements compared to randomly permuting SV locations (left bar).



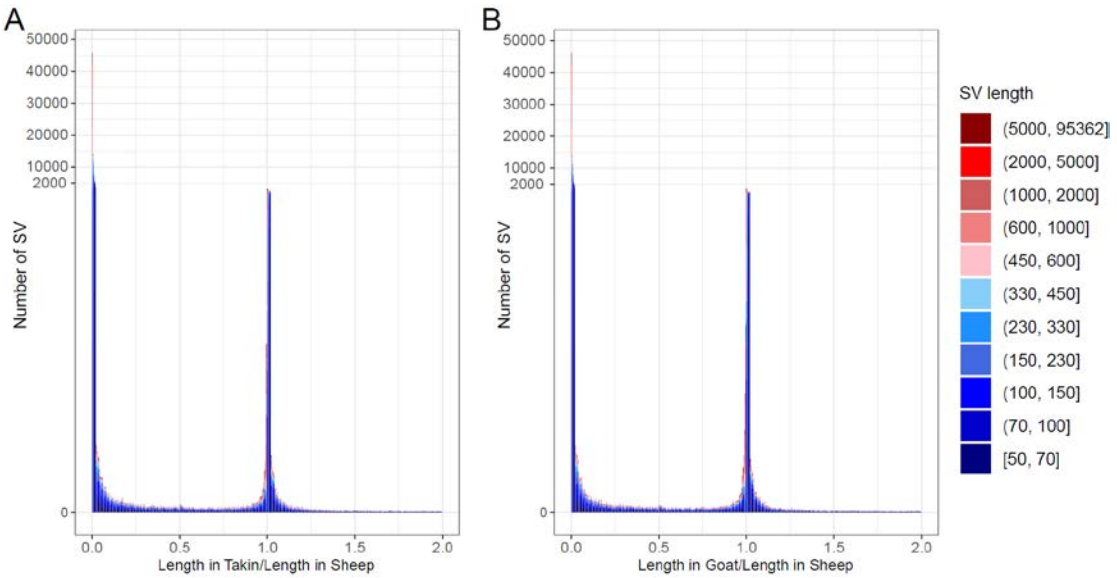
Supplemental Fig. S6 The SV hotspots are enriched in repetitive sequences, including TE (A), STR (B), Satellites (C) and VNTR (D). The enrichment analysis was performed using permutation test of 1000 times. The red dash lines represent the observed normalized number of SV hotspots intersecting with repeat sequences.



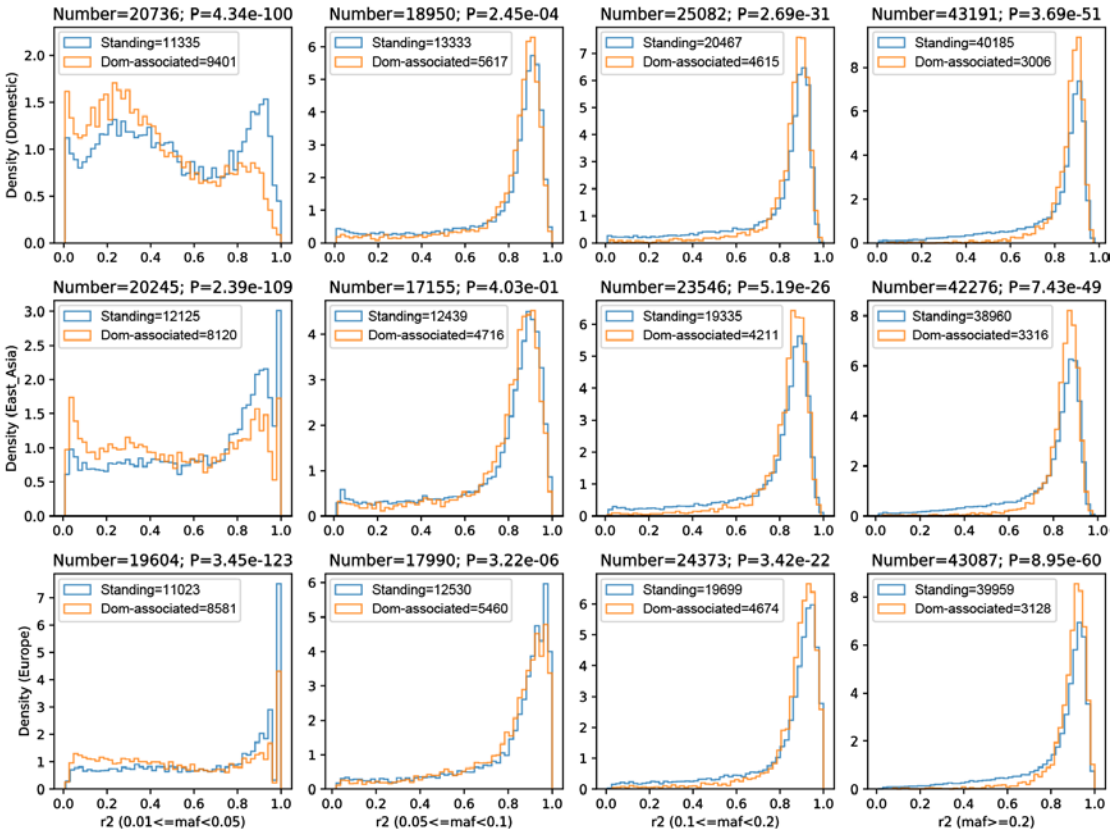
Supplemental Fig. S7 An ideogram showing SV hotspots and enrichment analysis of SVs with respect to the 5 Mb terminus of each chromosome. (A) The total number of SVs in each detected hotspot is shown by a scale going from blue to red. (B) The red dot represents the observed number of SVs at the 5 Mb terminus while the box plot shows the distribution of SV counts at the 5 Mb terminus, after 1000 random shuffling of SVs.



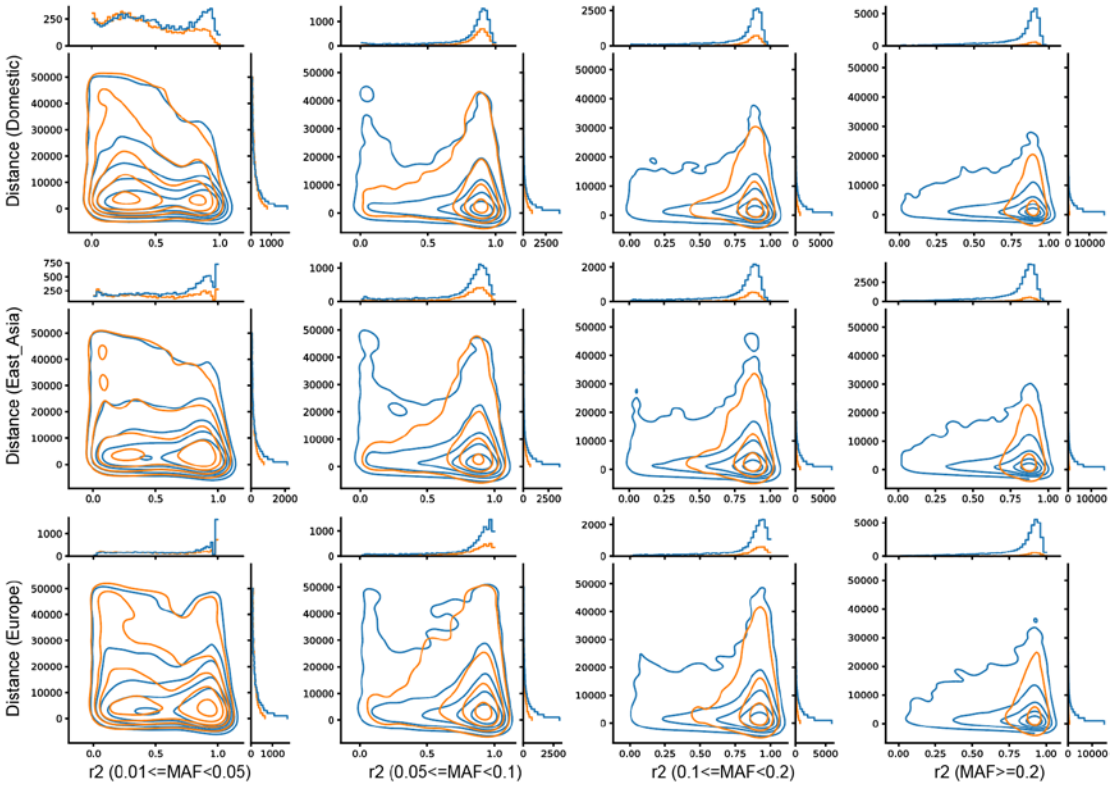
Supplemental Fig. S8 Segmental duplications identified in sheep genome and overrepresentation of SVs in SDs. (A) Circos plot highlighting intrachromosomal and interchromosomal SDs (>10 kb). (B) Box plot represents the distribution of SV counts in SDs after 1000 random shuffling of SVs, compared to the observed number of SVs in SDs (red dot).



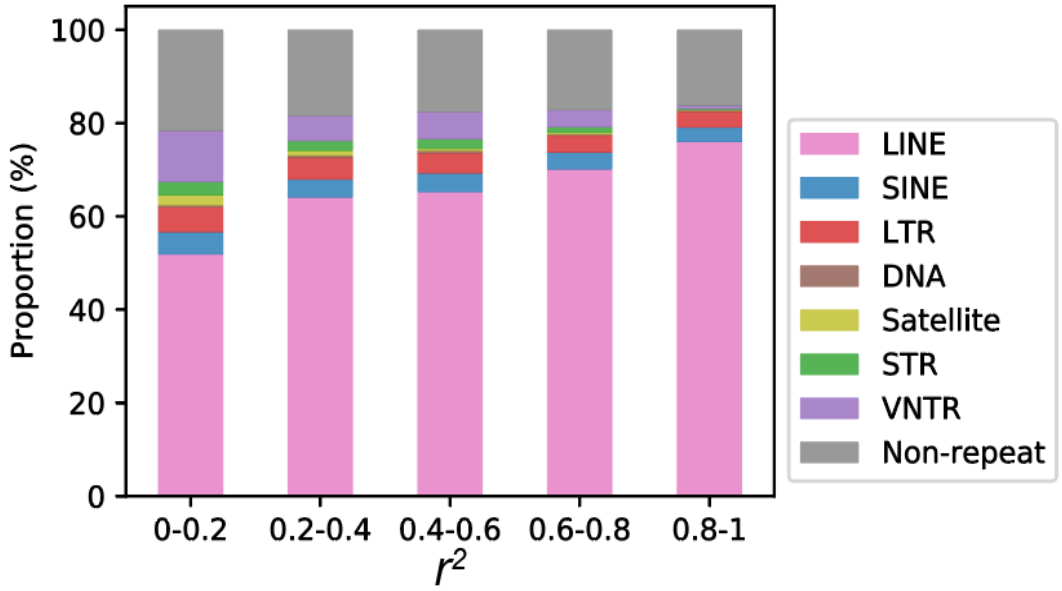
Supplemental Fig. S9 Ratio of SV sequence length in takin (A) and goat (B) to that in sheep. The SV sequences in sheep genome were lifted over to takin and goat genomes using whole genome alignment to determine whether the SV sequences were present in takin and goat. The two noticeable peaks at 0 and 1 represent absence and presence in the outgroup, respectively. Different SV size classes are indicated by different colors.



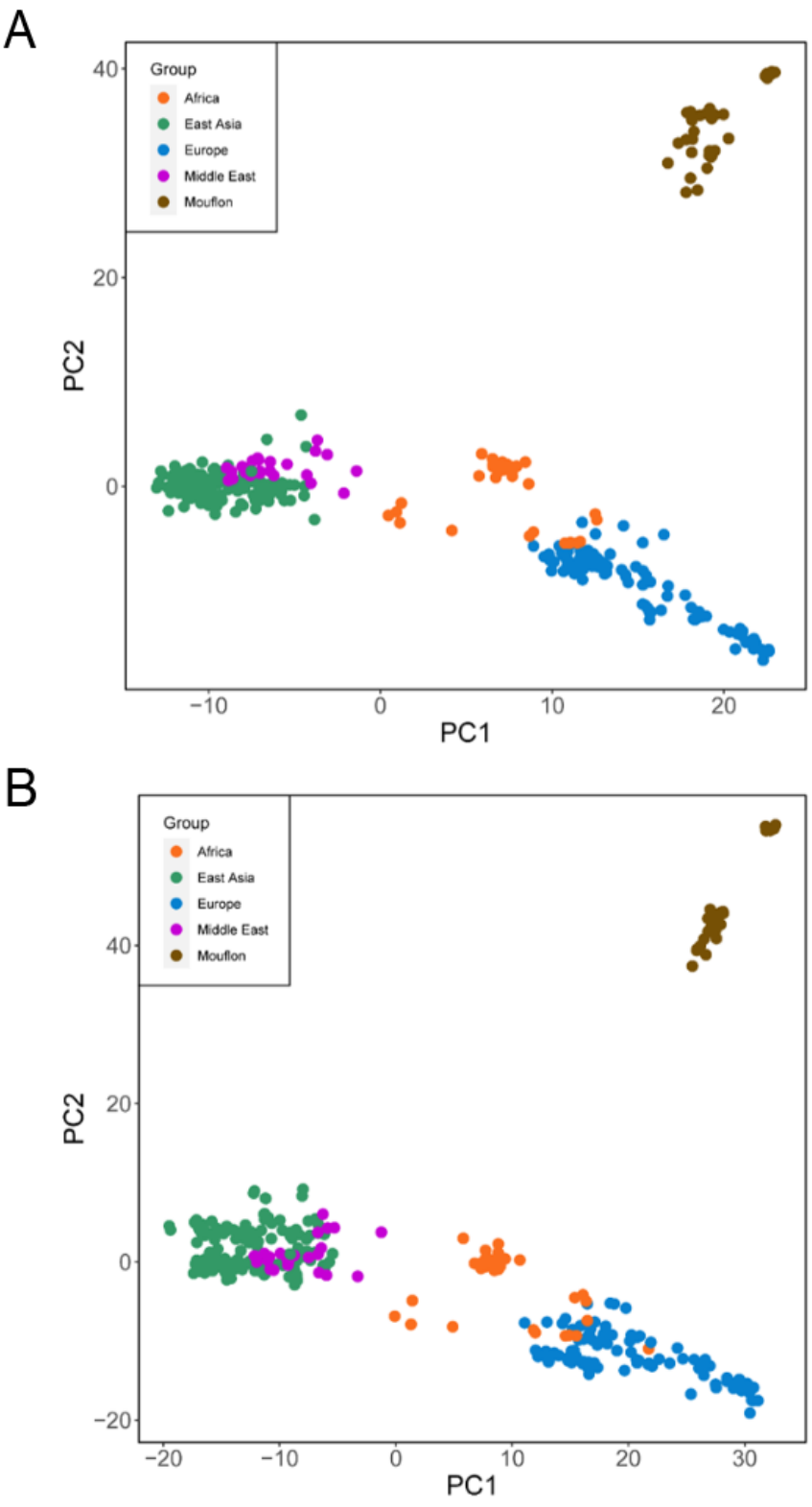
Supplemental Fig. S10 Density distribution of LD between SVs and nearby (±50 kb) SNPs in domestic, East Asian and European sheep, with different MAF ranges for SVs. Blue lines: Standing SVs that are present in Asiatic mouflons; Red lines: Domestication-associated SVs in sheep domestication.



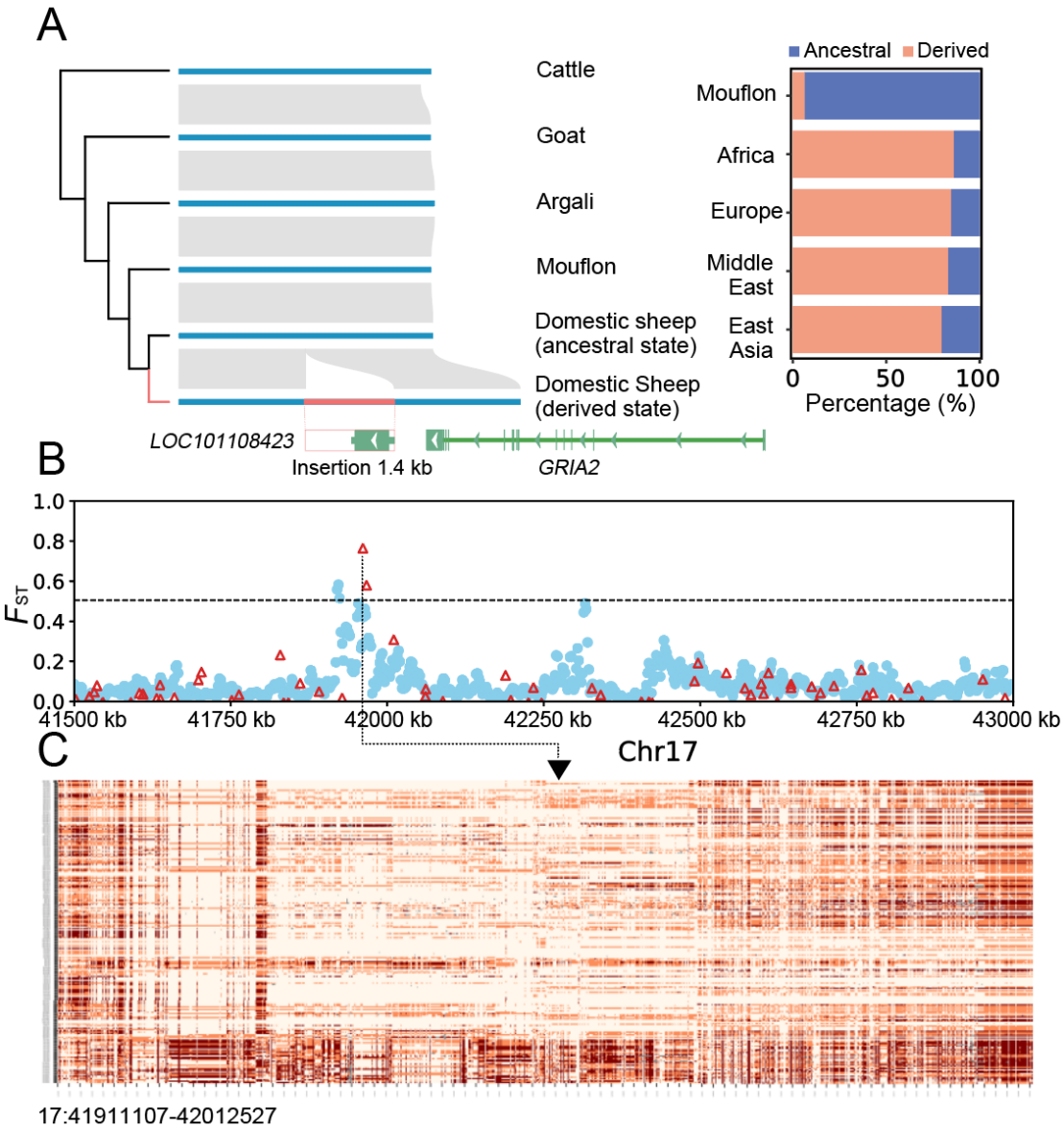
Supplemental Fig. S11 Contour density plots of LD between SVs and nearby (±50 kb) SNPs together with their physical distance in domestic, East Asian and European sheep, with different MAF ranges for SVs. Blue lines: Standing SVs that are present in Asiatic mouflons; Red lines: Domestication-associated SVs in sheep domestication.



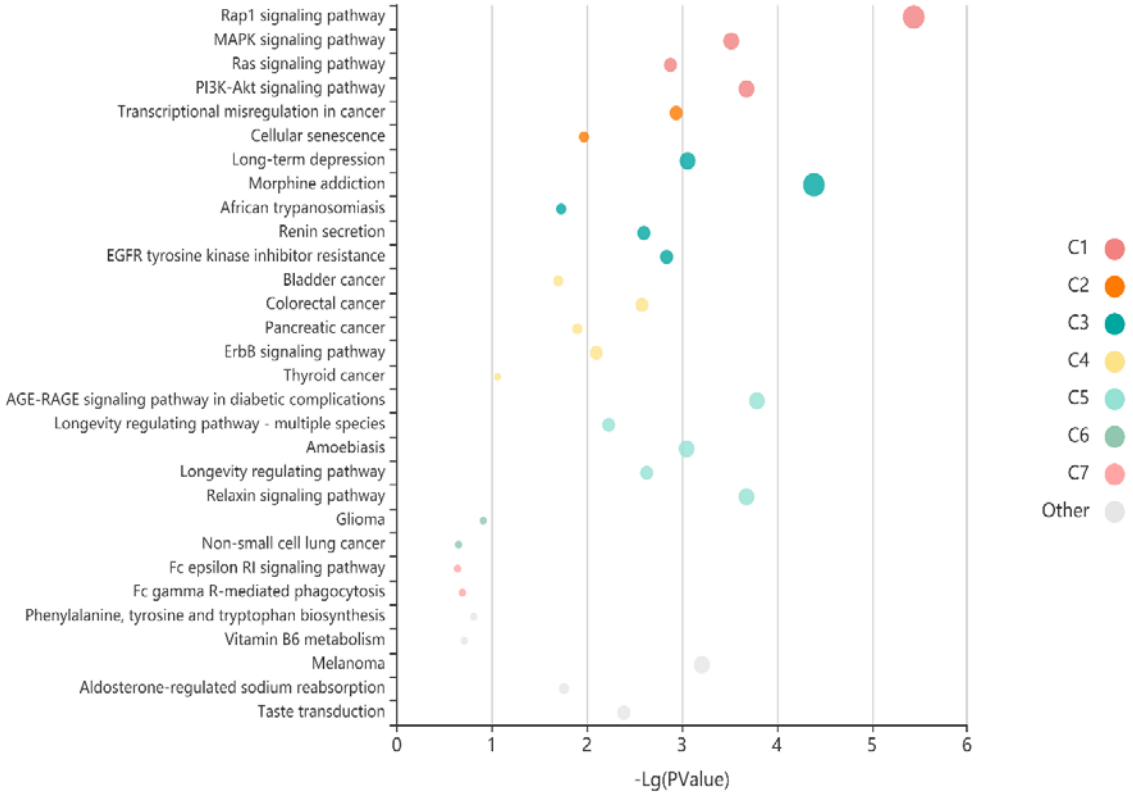
Supplemental Fig. S12 Repeat content of the SVs showing different LD with surrounding SNPs.



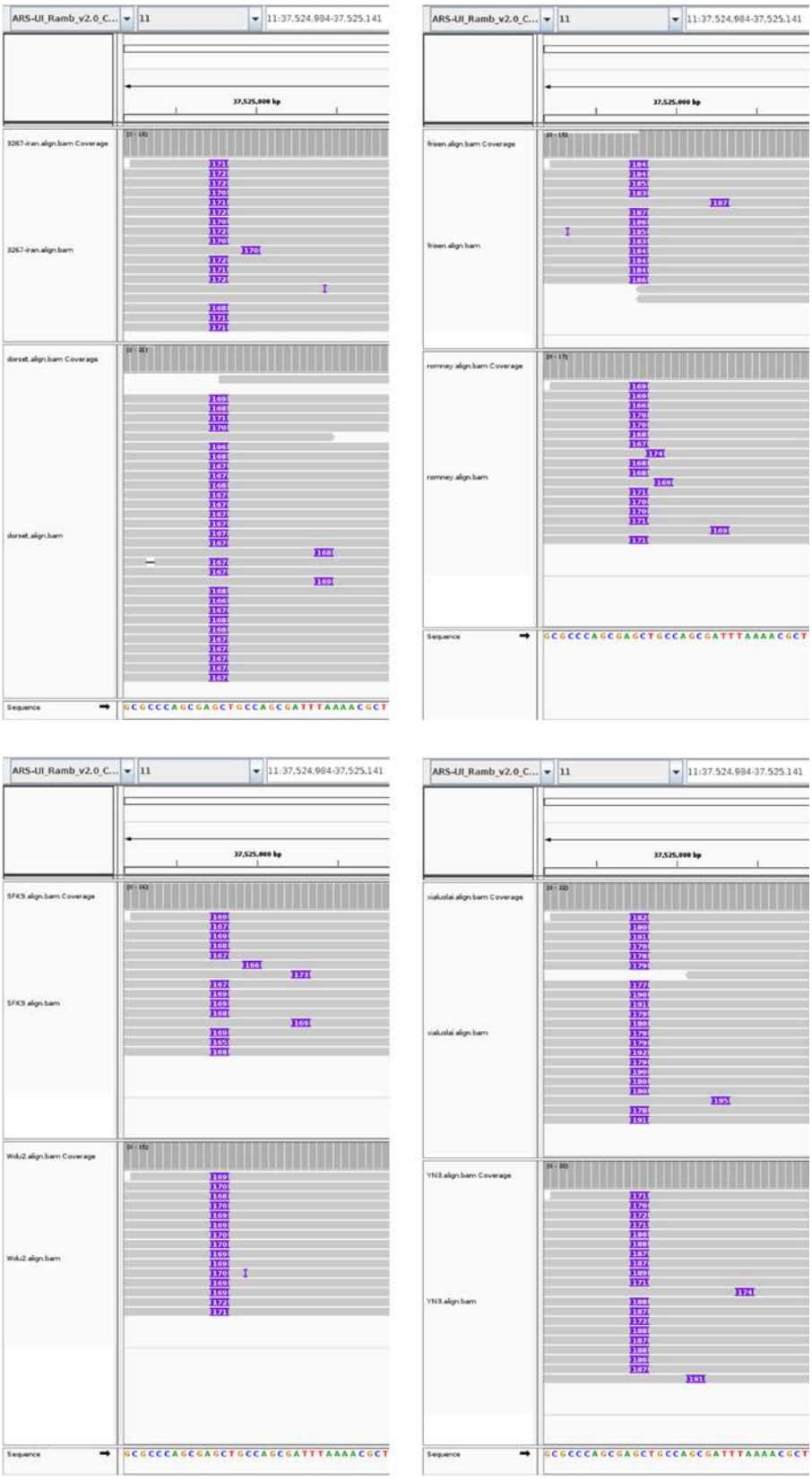
Supplemental Fig. S13 PCA plot showing the clusters of domestic sheep by insertions (A) and deletions (B).



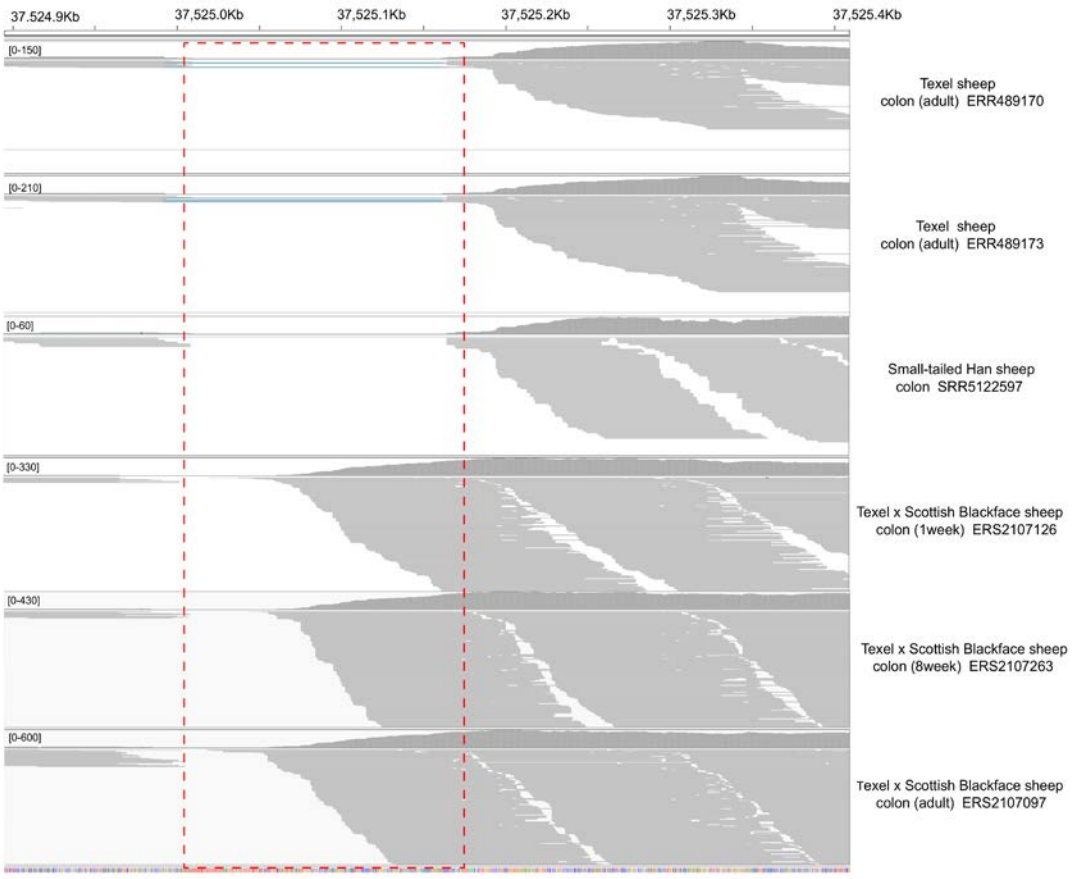
Supplemental Fig. S14 A 1.4 kb insertion near *GRIA2* gene displayed high differentiation between domestics and Asiatic mouflons. (A) The 1.4 kb insertion represents a derived state in domestic sheep, displaying high frequency in domestic sheep but rare in mouflons. (B) The F_{ST} of the 1.4 kb insertion and surrounding SNPs in 5kb window. SVs and SNPs were represented by red triangles and blue dots, respectively. (C) The haplotypes harboring the 1.4 kb insertion depicted by surrounding SNPs.



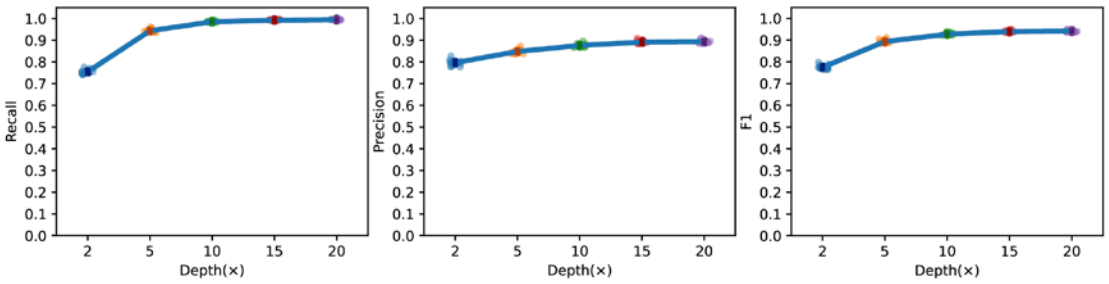
Supplemental Fig. S15 Enriched KEGG pathways for the population stratified SVs.



Supplemental Fig. S16 IGV profiles of the insertion adjacent to the annotated 5' UTR of *HOXB13*.



Supplemental Fig. S17 RNA-seq mapping for the 168 bp insertion shows that the insertion belongs to 5' UTR region. The insertion sequences were placed back into the reference genome highlighted by red box.



Supplemental Fig. S18 Evaluation of sequencing coverage of Illumina short reads on genotyping efficacy by measuring recall, precision and F1 score. The 15 individuals which we have both PacBio HiFi reads and Illumina reads (>20×) available are used. The Illumina reads were downsampled to 2×, 5×, 10×, 15×, 20× for Paragraph genotyping in compared with genotypes from PacBio HiFi reads.

Supplemental Tables

Supplemental Table S1 Sample list of the Illumina short-read data.

Supplemental Table S2 Information of the primary and partially phased assemblies.

Supplemental Table S3 Estimated sheep genome sizes from different sheep genomes in this study.

Supplemental Table S4 Centromeric and telomeric contents in unplaced contigs of the 15 primary assemblies.

Supplemental Table S5 List of the biallelic insertions and deletions identified in this study.

Supplemental Table S6 Information of divergent variations.

Supplemental Table S7 Information of multiallelic variations.

Supplemental Table S8 Manual validation of SVs. Fifty SVs were randomly selected from each range from 50-100bp, 100-500bp, 500-1000bp, 1000-5000bp and >5000bp.

Supplemental Table S9 List of the SDs identified in this study.

Supplemental Table S10 List of the SDs identified in this study with overlapped SVs.

Supplemental Table S11 Frequency of the derived SVs in the domestic sheep.

Supplemental Table S12 Frequency of the selected domestication-associated SVs in the domestic sheep.

Supplemental Table S13 SV frequency in 45 sheep breeds or populations

Supplemental Table S14 List of differentiated SVs between domestic sheep and Asiatic mouflon

Supplemental Table S15 KEGG analysis of domestication-stratified genes.

Supplemental Table S16 List of population-stratified SVs.

Supplemental Table S17 Accession numbers of the de novo assemblies.

Supplemental Data

SV sequences corresponding to SVs in Supplemental Table S5 were provided as a FASTA file with SV ID in the header of each FASTA record.

References

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573.

Chen N, Cai Y, Chen Q, Li R, Wang K, Huang Y, Hu S, Huang S, Zhang H, Zheng Z et al. 2018. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. *Nat Commun* **9**: 2337.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.

Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv*: arXiv:1308.2012.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AP et al. 2022. Segmental duplications and their variation in a complete human genome. *Science (80-)* **376**: eabj6965.